Collecting Primary Sources from Web Archives:

A Tale of Scarcity and Abundance

Federico Nanni

Data and Web Science Group

University of Mannheim

La diversité des témoignages historiques
est presque infinie.
(Bloch, 1949)

The World Wide Web is the largest collection of human testimonies that we have ever had at our fingertips. Spanning from institutional websites to digital libraries, from personal blogs to Twitter accounts of prominent politicians, from online newspapers to large-scale knowledge bases, an immense number of born-digital testimonies is waiting to be retrieved, selected and studied by future historians. In addition to this, while these new resources are piling up steadily in front of our eyes, they are also rapidly replacing their analogue counterparts, from printed news articles to personal diaries, from letter correspondences to scientific publications.

By acknowledging this sudden transition in production from printed to digital documents, the goal of this chapter is to present and discuss some of the new methodological issues that arise when these materials are to be employed as primary sources for studying the recent past. Firstly, an overview of the debate on the historian's craft is offered. Then, two different case studies that have dealt with the difficulties of adopting born-digital materials in historical work will be described: the first is focused on reconstructing the past of university websites as a new way for studying the recent past of academic institutions; the second retrieves materials from large-scale archives of the web in order to study contemporary socio-political events. Through these descriptions, it will be highlighted how a fruitful combination of the historical methods with approaches from other research areas, such as internet studies and natural language processing, could support future historians in successfully addressing them.

1. The Historical Method: Today and Tomorrow

In order to understand how the transition from analogue to digital sources is about to change the historian's craft, it is first of all essential to examine how the 'historical method' (Shafer, 1974) is generally defined and which are its major steps.

Defining a Subject

In the first part of any historical research, the scholar broadly defines the subject of investigation and - together with it - an initial question. The research question, firstly presented at a coarse-grained level, will be sharpened through the recursive process of collecting sources, interpreting them and by doing so discovering the underlying narrative.

Collecting the Evidence

In order to address the research question, the historian identifies the testimonies upon which she/he builds a narrative through a complex process of collection, analysis and selection of the remains of the past. These testimonies could be physical remains (e.g. buildings, statues), oral memories, printed documents (e.g. chronicles, diaries, articles, census data) and will soon become born-digital documents, such as websites, online forums, email threads, large-scale databases, etc. The process of collecting primary sources has been shaped and sharpened by decades of discussions in historiography both on how to establish the reliability of these materials, for example through source criticism, and on how much 'true knowledge' can be derived from them (there are as many interpretations of the same text as many readers, as Barthes (1967) has taught us).

Interpreting the Evidence

The interpretation of the collected textual sources represents the core of any historical research. Due to this reason, it has been the central focus of debate across 20th century historiography and has experienced drastic transitions in methodology. As a matter of fact, the analysis and interpretation of sources can be conducted in many different ways: traditional historiography scholarships have strongly relied on hermeneutics and on the careful qualitative examination of documents, while other approaches - which emerged during the second part of the 20th century inspired by social science methodologies (see the advent of Cliometrics - Greif, 1997) – have employed census data or economic reports in order to conduct large-scale quantitative analyses.

Through the '70s and the '80s postmodern and deconstructionist theories (starting from the works of Barthes, 1967; Derrida, 1967 and Lyotard, 1979, among others) have posed major critiques to the underlying assumption of both traditional and social-science historical scholarships that it is possible to discover a 'unique truth' about the past through to careful analysis of the remains. The enormous impact of these critiques has been remarked by many historians (Munslow, 2006; Burke, 2008) and has led to the so-called cultural turn in the profession, which is still reflected strongly today in the community[1].

Presenting a Narrative

The final step of any historical research is to define a narrative and write a history. The creation of a narrative, which is highly connected with the initial definition of the research question, gives the historian the possibility of placing the work she/he is writing as part of a larger contribution to the field. This is achieved in two interconnected ways: first of all, by offering a new/different perspective on the topic under study; in addition to this, by participating in the larger debate in historiography regarding the ways the past can be re-discovered, examined, described and - for certain authors[2] - even modelled.

1.1 A Computational Turn of the Craft?

History has been part of the so-called digital humanities (Schreibman et al., 2004), since their very beginning.[3] In particular, during the second part of the 20th century the potential of computational methods and their impact over the historian's craft have been recurrent topics in historiography. As Thomas III (2004) remarked, already in 1945 Vannevar Bush, in his famous essay ‘As We May Think’ pointed out that technology could be the solution that would enable us to manage the abundance of scientific and humanistic data (Bush, 1945); in his vision, the Memex could become an extremely useful instrument for historians.

The use of the computer in historical research, which grew significantly between the ‘60s and the ‘70s thanks both to the efforts of the Annales school (see for example Daumard and Furet, 1959) and to its application to the analysis of economic and census data (Greif, 1997), has been strongly related to the adoption of social science practices in historical studies (Evans, 2001). A pioneering work on the use of database technologies for historical research was conducted by Manfred Thaller during the ‘80s (Thaller, 1991).

However, as Milligan (2012) and Robertson (2016) have already remarked, a large majority of the historian community has remained skeptical towards the adoption of computational methods in the craft. This attitude has consolidated in opposition to other humanities disciplines: for example, in the last thirty years the field of literary study has largely experimented with the potential of what they have defined as ‘distant reading’ techniques, in order to extract quantifiable information from large amount of texts (Moretti, 2013). Instead, during the same time, the so-called digital history community (Cohen et al., 2008) has decided to focus primarily on the potentialities of the Web as a platform for the collection, presentation, and dissemination of material (Cohen and Rosenzweig, 2005) and on the more ‘communicative aspects’ of doing research in the humanities (Robertson, 2016). This can be noticed by observing the importance given to digital public history topics (Noiret, 2015), the relevance of teaching in digital history (Cohen et al., 2008) and the tradition of digital history mapping (Knowles and Hillier, 2008).

In the second part of the 2000s, thanks in particular to the prompt availability of digitized historical primary sources and the potentialities of web technologies, this skeptical attitude towards computational methods has slowly changed and a few interdisciplinary teams have developed tools in order to help other traditionally trained historians to employ these methods in their work. As Nelson (2016) remarked, the first fruitful applications of these methods for supporting historical narratives can be found in the works of Wilkens (2013) and Blevins (2014), which are robust examples of the beginning of a mature season of digital history.

While these early scholarships based on the use of computational approaches are essential for refreshing the historiographic debate, it is argued in this chapter that the adoption of computational methods could not be consider *per se* as a revolutionary turning point for the profession. In fact, use of these approaches is similar to other methodological turning points that historians have already experienced before (Milligan (2012), for example, identifies 'three waves' of computational history); moreover, during the last ten years the use of computational methods in humanities research has been strongly sustained and encouraged by public and private institutions (from the NEH Digital Humanities Advancement Grants to the Volkswagen Stiftung on 'Mixed Methods' in the Humanities) as well as private companies (e.g., Google's 'commitment' to the Digital Humanities) and often mainstream media sources (Rothman, 2014).

Nevertheless, it is argued in this chapter that historiography is about to experience a new and way more conspicuous turning point and that this will have a very strong impact on a specific step of the historian's craft, namely the way sources are collected from now on. Born-digital documents shared online, their ephemerality, preservation, availability and access is about to pose a large set of new challenges for future historians. In the next decades, the methodological debate in historiography will not only be centered around qualitative over quantitative, distant versus close, hermeneutics against statistical significance, but it will also address the needs of the community in finding ways of acquiring knowledge on our recent (digital) past.

1.2 The born-digital turn

The transition from analogue to born-digital materials is influencing the way historians study the past: materials such as websites, forums, blogs, tweets, emails, are in fact very different compared to traditional analogue and digitized primary sources. Born-digital materials have an extremely short life compared to printed documents as they are significantly more difficult to archive and preserve (LaFrance, 2015). This is due to a vast number of reasons (Brügger, 2005) and the consequence of it has been summarized by Rosenzweig (2003) with the concept of 'scarcity' of digital primary sources. Web pages disappear constantly from the live web (because they are removed by the author or by the owner of the platform, for instance due to copyright issues), leaving a familiar trace of 404 status code messages. Several scholars (Rosenzweig, 2003; Brügger, 2012 among others) have already remarked on the great impact that the ephemerality of web materials will have on the sharing and accessibility of the knowledge produced in the digital age for the next generations of historians. As it has been already said, in opposition to the fact that 'paper survives benign neglect for a long time' (Davis, 2014):

> The life cycle of most web pages runs its course in a matter of months. In 1997, the average lifespan of a web page was 44 days; in 2003, it was 100 days. Links go bad even faster. A 2008 analysis of links in 2,700 digital resources—the majority of which had no print counterpart—found that about 8 percent of links stopped working after one year. By 2011, when three years had passed, 30 percent of links in the collection were dead. (LaFrance, 2015)

Moreover, while some types of pages disappear more frequently than others (e.g. social media messages as opposed to official statements on administrative websites), those that do survive tend to change very frequently (Dougherty et al., 2010). For example, articles in newspapers (Nanni, 2013) as well as official administrative pages have been often modified without a specific mention (Owen and Davis, 2008). While initiatives such as the Internet Archive have a long tradition of preserving born-digital materials for future research, several issues still exist and new issues continue to emerge - not in the least due to constant innovations in web technologies. Therefore, researchers have to deal with the collected materials in a highly critical way, as Brügger (2012) described when he introduced his definition of web archive documents as reborn-digital materials:

> One of the main characteristics of web archiving is that the process of archiving itself may change what is archived, thus creating something that is not necessarily identical to what was once online. [...] And, second, that a website may be updated during the process of archiving, just as technical problems may occur whereby web elements which were initially online are not archived. Thus, it can be argued that the process of archiving creates the archived web on the basis of what was once online: the born-digital web material is reborn in the archive. (Brügger, 2012)

The difficulties in the preservation of digital sources present a new set of issues for historians who plan to employ them in their work; however, they remain only part of the overall problem. In fact, already in 2003, Rosenzweig envisioned that future historians will not only deal with a consistent scarcity of primary sources, but they will be also challenged by a never experienced before abundance of records of our past. The indispensable need of computational methods for processing and retrieving materials from these huge collections of primary sources has been a central topic of Milligan's publications (2012, 2016). From his works it emerges that now that the community is dealing with the abundance of born-digital sources, the use of computational approaches cannot be a choice for the digital humanities researcher anymore. Therefore, it becomes essential that the researchers adopt these solutions critically, always knowing their potential and limitations, and learn how to combine them fruitfully with the traditional historical method.

While the consequences of the advent of born-digital sources will be revolutionary for our profession, so far ‘very little attention has been paid to the new digital media as historical sources’ (Brügger, 2012), highlighting the fact that, while ‘new media is not that new anymore’ (Milligan, 2016) for our society, they remain a novelty for historians.

The next sections will remark further on this topic by describing two very different case studies that have dealt with the use of born-digital documents as primary sources for historical research. The first that will be introduced, focuses on examining the online presence of the University of Bologna, since the early Nineties, and remarks on the importance of combining the traditional historian’s craft with approaches from the field of internet studies.

2. Studying the Recent Past of Academic Institutions: A Tale of Scarcity

Multiple historians have considered academic institutions as political, economical and social actors; they have also argued how their power, role and influence changed over time, especially in relation to other actors, such as the city, the church, the national government (Brockliss, 1978). In particular, the comprehensive four-volume book series `*A History of the Universities in Europe’*, commissioned by the European University Association, edited by Hilde de Ridder-Symoens and Walter Rüegg and published between 1992 and 2011, offers an unprecedented overview on how universities have transformed over centuries: what they have taught and researched, how they have been institutionalized and how they have interacted with the society.

Historians of higher education, who presented their research in the volume, have adopted a large variety of primary and secondary sources in their works, from university-archive materials such as matriculation and graduation statistics to academic dissertations, from public reports to large scale statistical analyses. Based on these data, researchers have described and drawn conclusions on the history of universities on a large variety of topics, such as the way universities have managed resources, the way the admission process has changed before and after 1970, and how sciences and humanities have been taught and studied.

The current prompt availability of a large variety of born-digital materials such as syllabi (Cohen, 2005), bachelor, master and doctoral theses (Ramage, 2011), academic websites (Holzmann et al., 2016b) and their hyperlinked structure (Hale et al., 2014) is about to become a new relevant component of this field of research (Nanni, 2017b).

An emblematic example of the new challenges that born-digital documents will pose to historians of higher education is a study on reconstructing the recent past of the University of Bologna, through its digital sources (Nanni, 2017a).

The University of Bologna's website (Unibo.it), initially created in 1993, represents a new category of relevant resource for historians of higher education. The website collects and offers to the reader a large variety of documents, from descriptions of educational projects to overviews of research groups, from reports of collaboration with international institutions to information on opportunities of interactions with the private sector. In addition, it also shows how different departments, professors and research teams have been adopting the web – especially in its early days. Among the many relevant examples, one that deserves special mention is that the Astronomy Department of the university was already sharing preprints of their publications online in 1994 as html pages, in an early attempt of benefitting from the potential of the World Wide Web.

Nevertheless, while Unibo.it represents a useful collection of primary sources, the website has been modified several times during its first twenty years and the majority of the pages that have been published in the past are not available anymore on the live web. In particular, the transition to the so-called 'Portale D'Ateneo', which started in the early 2000s, required that all department pages change their structure and adopt a common layout and organization of their content. This has often forced the creation of brand-new department subdomains and the removal the previous versions of the same from the live web. As an additional issue, the team that has managed the website during this entire transition has not consistently archived the previous versions of the website and documented their work.

Given the fact that as of 2017 the National Libraries of Florence and Rome are still not part of the International Internet Preservation Consortium (IIPC) and no coordinated project with the specific purpose of preserving the national websphere currently exists in Italy, the Internet Archive remains the only resource available for recollecting all the materials that are not available on the University of Bologna website anymore. However, in 2002 a removal request[4] from the administrative team of Unibo.it was sent to the Internet Archive, and for this reason Unibo.it had been inaccessible through the Wayback Machine for more than thirteen years. This highly complex situation reflects a new level of difficulties that future historians will encounter while attempting to collect born-digital sources. In the next section, an overview of the variety of sources and methods that have been used to deal with this issue and to reconstruct the past of Unibo.it will be presented.

Library and Archive Materials

As an initial step of the research, materials available in the university library and archives were consulted. Among many other documents, a very useful source has been the university yearbook. In the early 90s only a few pieces of information regarding the website were mentioned in the yearbook; nevertheless, this source offered an initial diachronic overview of the official teams that were managing Unibo.it and was useful for drawing a list of people to interview.

Interviews

In order to capture the rationale and the changing architecture of the website, the different teams who managed the website were interviewed, together with technicians and researchers who worked on the development of the pages of various departments, especially during the '90s. Yet another interesting finding, presumably highly relevant for future historians, was that many times during the interviews the subjects used public and private backups of emails in order to recollect the memories of their experience in working on Unibo.it and to confirm passages of the historical reconstruction.

Newspapers

As already done in previous work (Brügger, 2011), where printed media were used to retrieve information about the web of the past, information related to Unibo.it and the role of the website for the University of Bologna have been identified in local and national newspaper archives. During the ‘90s, newspapers such as *La Repubblica* and *Il Resto del Carlino* published a few short articles covering the new functionalities on the website (e.g. free email account for all students, online fee payments, etc.). These publications, together with materials collected from the university digital magazines (*Alma2000*, *AlmaNews*, *Unibo Magazine*), offered an additional overview on how the university decided to promote the website to its audience.

Online Forums

To get a closer look at the everyday use of the website by students and researchers, other materials have been collected and analyzed, starting from student forums (e.g. *UniversiBo*) and Usenet discussions preserved by Google. These documents, especially in the ‘90s, present the perspective and enthusiasm of a rather small but specific subset of the university community, namely students, researchers and professors in STEM fields, whose departments were among the first ones to offer access to the web.

Live Web Materials

While the website has been restructured multiple times during its first 20 years online, many resources are still available on the live web and can reveal the current role of website in the university's organization and management (e.g. attracting national and international students and researches, promoting collaborations with the private sector, etc). Additionally, the social media pages of the institution (such as Facebook, Youtube and Twitter profiles) are becoming key components of its presence online, showing alternative and more informal ways of interaction with the users.

Presence of Italian Websites in Other National Web Archives

Aside from the Internet Archive, since 1996 national libraries from all around the world have also begun to preserve their national web past. PANDORA, started in 1996 by the National Library of Australia, the UK Web Archive (2004), the Netarkivet (2005) in Denmark and the Portuguese Web Archive (2011) are just a few examples of this international endeavor. Given the complexity of defining and preserving what is called a 'national web-sphere' (Brügger, 2009), this research also explored the use of foreign web archives as a proxy for studying Unibo.it. The practice of retrieving primary sources related to an Italian university website in foreign web archives could seem rather odd as the goal of a national web archive is precisely to preserve the web of its country, however from time to time part of the non-national web also ends up being preserved, unintentionally, by these digital archives.

For example, to archive national web spheres in an automatic way, archivists could set up crawlers with a maximum number of hyperlinks they can follow, with a specific set of starting points. A crawler which is set to go at most ten links away from one of these URLs could also end up crawling non-national content, as it will systematically follow all the hyperlinks. For this reason, if the University of Bologna were to organize a Summer School and Aarhus University had linked it from its website, the University of Bologna website (or at least part of it) would be unintentionally preserved in the Danish Web Archive.

As a part of this work, it has been found out that both the Portuguese (Arquivo) and Danish (Netarkivet) web archives have preserved parts of Unibo.it several times, since 2006.

Cloned Versions of the Website

Among the variety of sources available, one deserves a specific mention. In May 2007, a group of activists decided to create a copy of the Unibo.it web interface, as part of a protest against the European Credit Transfer and Accumulation System (ECTS) for the evaluation of the number of hours of study. In the URL *http://www.unibologna.eu* an identical version of the website was available, with the description of the reasons of the protest.

This source has not only been important in this study as it documented an innovative way of conducting a protest against an academic institution (by targeting its website), but also because the cloned-website was preserved by the Internet Archive.

A Critical Combination of Sources and Methods

The combination of traditional archival practices with approaches from the field of internet studies is essential in the attempt of facing this emblematic example of scarcity of born-digital primary sources and reconstructing the past of the University of Bologna website. This new methodology for collecting born-digital evidences has been especially useful in identifying the narrative behind the early years of Unibo.it, which involves the arrival of a Turkish professor from the United States at the university in 1988, the establishment of the second Italian node to the Internet and the creation of arguably one of the most relevant university websites of the country[5].

While the difficulties in reconstructing the recent past of a university website could surprise the reader, as less than 30 years have been passed since its creation, they only represent one part of the new issues that born-digital sources will pose to future historian.

As it has been previously remarked and will be expanded in the next section, future historians will be in fact also challenged by a never experienced before abundance of records of our past. The second case study presented in this chapter focuses on obtaining small topic-specific collections from large-scale archives of the web; by presenting the encountered challenges and describing the adopted solutions, it will be remarked on the importance of fruitfully combining the traditional historical method with approaches from the field of natural language processing.

3. Creating Political Event Collections: A Tale of Abundance

The World Wide Web provides the research community with an unprecedented abundance of primary sources for diachronically tracing, examining and understanding major events and transformations in our society. For two decades, public and private institutions have preserved these born-digital materials for future analysis (Gomes and Costa, 2011). However, these collections are now so large that – in the rare cases when they are fully available for research (Hockx-Yu, 2014) – it is not feasible for scholars to study political and social phenomena by examining them in their entirety. If we for instance consider the Internet Archive, during its first twenty years it has preserved almost 500 billion web pages, and as of 2017 it has a collection of around 25 petabytes of data. Since 2001, this collection has become available for research through a URL search tool on the Wayback Machine. In the most recent years, information retrieval systems supporting keyword search over the diachronic layers of web archives have been developed by the research community and employed by institutions such as the UK Web Archive and – since 2017 – also partially by the Internet Archive. In addition to this, out-of-the-box tools such as ArchiveSpark (Holzmann et al., 2016a) and Warcbase (Lin et al., 2017) have been developed by the research community with the specific goal of supporting scholars in gathering information from large-scale web archive collections.

One of the main endeavors of web archive institutions for fostering the use of these new resources is to offer manually curated sub-collections regarding recent socio-political events. On Archive-It – a subscription web archiving service provided by the Internet Archive – a few collections regarding large-scale events such as the Boston Marathon Shooting, the Black Lives Matter movement and the Charlie Hebdo terrorist attack are available. The collections are curated by 'the Archive-It team in conjunction with curators and subject matter experts from institutions around the world'.

In addition to manual selection, another solution employed by digital archivists for creating and sharing these event collections is to adopt a filtering approach that presents to the user only those documents that mention the name of the event. This type of approach is common in event-harvesting from Twitter, where researchers collect all tweets that – for example – mention the hashtag of the event.

While both collecting documents from web archives through manual selection and retrieving materials through name-filtering have already proved their usefulness in supporting researchers in the humanities and social sciences (e.g., Small, 2011), they have a few crucial limitations. On one hand, manual selection is obviously a painstakingly long process – given the previously mentioned difficulties of retrieving information from web archives. On the other hand, collecting documents using the event-name heuristics presents the crucial limitation of often missing information on background stories as well as premises of the examined events. To give a specific example, let us imagine that the goal is to collect primary sources regarding the 2004 Ukraine Orange Revolution. If the adopted method only retrieves documents that mention the name of the event, it will not collect materials that connect the premises of the revolution to the previous controversial presidential election in the country. And the same issue will emerge when studying the first free Algerian elections since their independence (1990), which is a premise of the Algerian civil war, or even when investigating the economic crisis behind Fujimori's auto-golpe in Peru, 1992. In this last case, the documents that discuss the adoption of austerity measures will be not be part of the collection. Moreover, the name used for referring to an event might change over time or vary between countries and languages: for example, one of the early hashtags used for the 2011 Egyptian Revolution was #jan25, referring to the day it started.

The second case study presented in this chapter is an interdisciplinary project between computer science and political history focused on building more comprehensive sub-collections regarding events such as elections, protests and political crises from large-scale web archives. As part of this research, a system that employs natural language processing methods and information retrieval approaches has been developed, which is able to gather and organize a highly comprehensive collection of sources describing a specific event (Nanni et al., 2017). The developed approach is inspired by the fact that, when historians are conducting the same task manually (i.e., identifying relevant materials across an entire archive), they do not necessarily search only for documents that mention the name of the event. What historians will try to collect are also those documents that talk about related aspects which provide the context, involving for example some of the participants to the event, but not others. If we consider the previous example regarding the Orange Revolution, historians will also be interested in materials from the same period of time discussing the political career of Yulia Tymoshenko or addressing the state of the political relations between Ukraine, Russia and the European Union.

Identifying Related Concepts and Entities

In order to achieve this goal in an automated fashion, the first step is to be able to identify a set of concepts and entities that are relevant to an event. To do so, DBpedia (Auer et al., 2007) has been employed. This is a large-scale knowledge base extracted from Wikipedia, where events (such as the

Orange Revolution) are represented by nodes and connected through edges (i.e., hyperlinks in Wikipedia) to other related entities.

Retrieving Contextual Passages

For each collected entity and concept, a textual passage presenting it in the context of the event was also extracted from Wikipedia (for example: 'Yulia Tymoshenko co-led the Orange Revolution and was the first woman appointed Prime Minister of Ukraine'). This is an optimal solution for identifying other terms that could be useful to indentify relevant documents.

Ranking Concepts and Entities

Having obtained an initial set of potentially relevant concepts and entities, the goal is to score each of them on how relevant they are to the event. For example, while Yulia Tymoshenko is highly relevant for the Orange Revolution, the European Union played only a marginal role in the event. Different approaches for ranking entities and concepts for relevance were tested and the best performing solution was to compute distances between entities and the event employing out-of-the-box RDF vector-representations (Ristoski and Paulheim, 2016).

Finding Mentions in Text

Having our ranked set of entities and concepts, other documents were retrieved from the web-archive mentioning them in relevant contexts. In order to go beyond simple string-matching of concepts that are considered relevant (e.g., 'protests', 'revolution', 'crisis', 'election'), word-embedding representations (Mikolov et al., 2013) have been adopted. Embedding techniques represent each word, entity or concept (e.g., 'protest') as a numeric-vector of *n* dimensions. This allows to measure similarity across different words and to collect relevant materials even if they talk about 'demonstration' or 'crisis', instead for example of mentioning 'protest' or 'revolution'.

Final Collection Building

It could happen that documents mention relevant entities and concepts out of context, for example as part of a comparison: 'The popular opposition to Ethiopia's current corrupt regime is comparable to the Orange Revolution in Ukraine.’. In order to filter them out and select only the documents that should be included in the event-collection, a machine learning system called Learning to Rank (Liu, 2009) has been employed, which, given an initial set of relevant and not relevant documents, learns how to abstract this property and to automate the ranking process.

A Critical Combination of Sources and Methods

The combination of traditional practices of historical research with methodologies and approaches from the fields of natural language processing and information retrieval is essential for facing the large abundance of born-digital primary sources. Some of the approaches presented in this chapter have been already adopted in political science research. One of these first studies focuses on retrieving documents which referred to political events (e.g., elections) from institutional web collections of the United States government in order to define a new measure of ‘attention’ of the U.S. Congress and the President to democratization and electoral practices in other countries, from Zimbabwe to Haiti and Egypt (Elshehawy et al., 2017). By doing so, this initial work highlights both the potential and challenges of using born-digital documents and computational methods for obtaining new insights on the recent political past.

The two case studies presented in this chapter reveal the importance of adopting a highly interdisciplinary approach when dealing with born-digital sources; methodologies from the field of internet studies could support historians in reconstructing lost web pages, while natural language processing methods could guide them in retrieving documents from large-scale web archives. The final part of this chapter will remark further on this, by discussing on the importance of offering this interdisciplinary preparation to future historians in their educational programs.

4. Conclusion: A New Generation of Historians

In recent years, researchers have argued that history, as other humanities disciplines, is reaching a turning point in its methodology (Scheinfeldt, 2012; Graham, Milligan and Weingart, 2015; Nelson, 2016): sustained by the efforts of many digitization projects, the community has been employing computational methods in order to examine these vast resources and obtaining new insights. This change in methodology has reopened a long-term debate regarding the ways textual evidence of the past can and shall be properly interpreted.

While for the historical profession it is of course beneficial to constantly debate and criticize the validity of established practices of acquiring knowledge from sources, it is argued in this chapter that the adoption of digitized datasets and computational methods cannot be considered, by itself, the triggering factor of a fundamental turning point in our profession. In fact, adopting (or not) large-scale datasets of digitized sources, together with computational methods, will always remain a choice for the history scholar: Charles Darwin can still be studied without conducting text mining over the collections presented on Darwin Online, as well as the London of 18th century can be examined without distant reading the Proceedings of the Old Bailey Online.

However, it is also argued that history is in fact about to face a paradigm-shifting transition in its methods, but the triggering cause of this transition relies on the born-digital nature of the large majority of sources produced by contemporary societies. This change affects any type of document we create and consume in our everyday life, from bureaucratic forms collected by the public sector to newspapers articles to political mail correspondences to university websites, and it is about to present its multifarious consequences on historical research.

Born-digital sources are significantly more complex to archive, collect, analyze and select compared to traditional materials. Websites (such as Unibo.it), are large and variegated collections of documents, which are often not preserved in their entirety by web archive initiatives and can be re-constructed only through the meticulous combination of various pieces of information from different sources. When a resource, such as the institutional website of an administration is finally re-created, it is often so vast that computational technologies (i.e. natural language processing methods and information retrieval approaches) are necessary for identifying and retrieving specific documents.

The methodological steps overviewed in this chapter for collecting, analyzing and selecting born-digital documents require strong interdisciplinary competences and a highly critical attitude towards sources and methods. In this complex scenario, this chapter concludes by raising a very pressing question: how can the new generations of historians be prepared to face these new challenges?

In recent years, the digital history community has already offered many educational activities on computational methods to its students. From workshops to panels, from courses to summer schools, from tutorials to hackathons, these initiatives have almost always been focused on presenting the potential of new resources, tools and platforms to the history students, following an attitude which has been branded as 'more hack, less yack' (Nowviskie, 2014). While offering hands-on experiences with computational tools is important in order to introduce history students to the digital humanities, a critical approach is strongly needed in order to properly deal with born-digital sources and computational methods.

For this reason, it is essential that students will first of all be guided in shaping their research topics and receive early on in their studies the preparation necessary to support a critical analysis of the born-digital documents and computational methods at their disposal. This will be imperative for a generation of historians who will be able to go beyond an unquestioned adoption of the new sources and tools at their disposal and will instead critically employ them, in search of new historical perspectives.

References

Auer S., Bizer C., Kobilarov G., Lehmann J., Cyganiak R., Ives Z. (2007) 'DBpedia: A Nucleus for a Web of Open Data', *Proceedings of the 6th International and 2nd Asian Conference on Semantic Web:* 722-735.

Barthes, R. (1967) 'Discourse on History', *Social Science Information*, *6*(4): 65-75.

Blevins, C. (2014) 'Space, Nation, and the Triumph of Region: A View of the World from Houston', *The Journal of American History*, *101*(1): 122-147.

Bloch, M. (1949*) Apologie pour l'histoire, ou, Métier d'historien*, Armand Colin, Paris.

Brockliss, L. W. (1978) 'Patterns of Attendance at the University of Paris, 1400–1800', *The Historical Journal*, 21(3): 503-544.

Brügger, N. (2005) 'Archiving Websites: General Considerations and Strategies', *The Centre for Internet Research*, Aarhus.

Brügger, N. (2009) 'Website History and the Website as an Object of Study'. *New Media & Society*, *11*(1-2): 115-132.

Brügger, N. (2011) 'Web Archiving–between Past, Present, and Future', in M. Consalvo and C. Ess (ed.), *The Handbook of Internet Studies*, Wiley-Blackwell, Oxford.

Brügger, N. (2012) 'When the Present Web is Later the Past: Web Historiography, Digital History, and Internet Studies'. *Historical Social Research/Historische Sozialforschung*, 37(4): 102-117.

Burke, P. (2008) *What is Cultural History?*, Polity, Cambridge (UK).

Bush, V. (1945) ‘As We May Think’, *The Atlantic Monthly*, *176*(1): 101-108.

Cohen, D. J., & Rosenzweig, R. (2005) *Digital History: A Guide to Gathering, Preserving, and Presenting the Past on the Web*, University of Pennsylvania Press.

Cohen, D. J. (2005) ‘By the book: Assessing the Place of Textbooks in US Survey Courses’, *The Journal of American History*, 91(4): 1405-1415.

Cohen, D. J., Frisch, M., Gallagher, P., Mintz, S., Sword, K., Taylor, A. M., & Turkel, W. J. (2008) ‘Interchange: The Promise of Digital History’, *The Journal of American History*, 95(2): 452-491.

Daumard, A., & Furet, F. (1959) ‘Méthodes de l'histoire sociale: les archives notariales et la mécanographie’. *Annales. Histoire, Sciences Sociales*, 14 (4): 676-693.

Davis, C. (2014) ‘Archiving the Web: A Case Study from the University of Victoria’. *code {4} lib Journal*, 26 (http://journal.code4lib.org/articles/10015)

Derrida, J. (1967) *Of Grammatology*, Les Éditions de Minuit, Paris.

Dougherty, M., Meyer, E. T., Madsen, C. M., Van den Heuvel, C., Thomas, A., & Wyatt, S. (2010), ‘Researcher Engagement with Web Archives: State of the Art’, *Preprint on SSRN* (https://ssrn.com/abstract=1714997)

Elshehawy, A., Marinov, N., & Nanni, F. (2017) ‘Quantifying Attention to Foreign Elections with Text Analysis of US Congress and the Presidency’, *Preprint on SSRN* (https://ssrn.com/abstract=2981486)

Evans, R. J. (2001) *In Defence of History*, Granta Books, London.

Fogel, R. W., & Engerman, S. L. (1974) *Time on the Cross*, University Press of America, Lanham, Maryland.

Gomes D., Miranda J., Costa M. (2011) 'A Survey on Web Archiving Initiatives', *Proceedings of the 15th international conference on Theory and practice of digital libraries*: 408-420.

Graham, S., Milligan, I., & Weingart, S. (2015) *Exploring Big Historical Data: The Historian's Macroscope*, Imperial College Press, London.

Greif, A. (1997) 'Cliometrics After 40 Years', *The American Economic Review*, *87*(2): 400-403.

Hale, S. A., Yasseri, T., Cowls, J., Meyer, E. T., Schroeder, R., & Margetts, H. (2014) 'Mapping the UK Webspace: Fifteen Years of British Universities on the Web', *Proceedings of the 2014 ACM Conference on Web Science*: 62-70.

Hockx-Yu, H. (2014) 'Access and Scholarly Use of Web Archives', *Alexandria: The Journal of National and International Library and Information Issues*, 25(1-2): 113-127.

Holzmann, H., Goel, V., & Anand, A. (2016a) 'Archivespark: Efficient Web Archive Access, Extraction and Derivation', *Proceedings of* the *2016 IEEE/ACM Joint Conference on Digital Libraries (JCDL)*: 83-92.

Holzmann, H., Nejdl, W., & Anand, A. (2016b) 'The Dawn of Today's Popular Domains: A Study of the Archived German Web Over 18 Years', *Proceedings of* the *2016 IEEE/ACM Joint Conference on Digital Libraries (JCDL)*: 73-82.

Iggers, G. G. (2005) *Historiography in the Twentieth Century: From Scientific Objectivity to the Postmodern Challenge*, Wesleyan University Press, Middletown (CT).

Knowles, A. K., & Hillier, A. (eds) (2008) *Placing History: How Maps, Spatial Data, and GIS are Changing Historical Scholarship*, ESRI, New York.

LaFrance, A. (2015) 'Raiders of the Lost Web'. *The Atlantic*, 14 (https://www.theatlantic.com/technology/archive/2015/10/raiders-of-the-lost-web/409210/)

Lin, J., Milligan, I., Wiebe, J., & Zhou, A. (2017) 'Warcbase: Scalable Analytics Infrastructure for Exploring Web Archives', *Journal on Computing and Cultural Heritage (JOCCH)*, 10(4): 22.

Liu, T. Y. (2009) 'Learning to Rank for Information Retrieval', *Foundations and Trends in Information Retrieval*, 3(3): 225-331.

Lyotard, J. F. (1979) *The postmodern condition: A Report on Knowledge*, Minuit, Paris.

Munslow, A. (2006) *Deconstructing History*. Routledge, New York.

Milligan, I. (2012) 'Mining the 'Internet Graveyard': Rethinking the Historians' Toolkit'. *Journal of the Canadian Historical Association/Revue de la Société historique du Canada*, 23(2): 21-64.

Milligan, I. (2016) 'Lost in the Infinite Archive: The Promise and Pitfalls of Web Archives', *International Journal of Humanities and Arts Computing*, *10*(1): 78-94.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013) 'Distributed Representations of Words and Phrases and their Compositionality', *Proceedings of the 26th International Conference on Neural Information Processing Systems*: 3111-3119.

Moretti, F. (2013) *Distant Reading*. Verso Books, London.

Nanni, F. (2013) ‘L’archiviazione delle pagine dei quotidiani online’, *Diacronie. Studi di Storia Contemporanea*, 15 (3) (http://www.studistorici.com/wp-content/uploads/2013/10/02_NANNI.pdf)

Nanni, F. (2017a) ‘Reconstructing a Website's Lost Past: Methodological Issues Concerning the History of www.unibo.it’, *Digital Humanities Quarterly*. 11(2) (http://www.digitalhumanities.org/dhq/vol/11/2/000292/000292.html)

Nanni, F. (2017b) ‘The Web as a Historical Corpus: Collecting, Analysing and Selecting Sources on the Recent Past of Academic Institutions’, Ph.D. Dissertation, University of Bologna.

Nanni, F., Ponzetto, S. P., & Dietz, L. (2017) ‘Building Entity-Centric Event Collections’, *Proceedings of 2017 IEEE/ACM Joint Conference on Digital Libraries (JCDL)*: 199-209.

Nelson, R. K. (2016) ‘Digital Humanities as Appendix’, *American Quarterly*, 68(1): 131-136.

Noiret, S. (2015) ‘Digital Public History: Bringing the Public Back In’, *Public History Weekly*, *3(13) (http://hdl.handle.net/1814/38393)*.

Nowviskie, B. (2014) ‘On the Origin of “Hack” and “Yack”’, , in M. K. Gold and L. F. Klein (eds) *Debates in Digital Humanities (2nd edn)*, University of Minnesota Press (http://dhdebates.gc.cuny.edu/debates/text/58)

Owen, D., & Davis, R. (2008) ‘Presidential Communication in the Internet Era’, *Presidential Studies Quarterly*, *38*(4): 658-673.

Ramage, D. R. (2011) ‘Studying People, Organizations, and the Web with Statistical Text Models’, Ph.D. Dissertation, Stanford University.

Ristoski, P., & Paulheim, H. (2016) 'RDF2vec: RDF Graph Embeddings for Data Mining', *Proceedings of the 2016 International Semantic Web Conference*: 498-514.

Robertson, S. (2016) 'The Differences Between Digital History and Digital Humanities', in M. K. Gold and L. F. Klein (eds) *Debates in Digital Humanities (2nd edn)*, University of Minnesota Press. (http://dhdebates.gc.cuny.edu/debates/text/76).

Rosenzweig, R. (2003) 'Scarcity or Abundance? Preserving the Past in a Digital Era', *The American Historical Review*, 108(3): 735-762.

Rothman, J. (2014) 'An Attempt to Discover the Laws of Literature', The New Yorker.

Rüegg, W., & de Ridder-Symoens, H. (eds) (1992) *A History of the University in Europe*, Cambridge University Press, Cambridge.

Scheinfeldt, T. (2012) 'Sunset for Ideology, Sunrise for Methodology', in M. K. Gold and L. F. Klein (eds) *Debates in Digital Humanities (1st edn)*, University of Minnesota Press: 124-127.

Schreibman, S., Siemens, R., & Unsworth, J. (eds) (2004) *A Companion to Digital Humanities*, Blackwell Publishing, Oxford.

Shafer, R. J. (1974) *A Guide to Historical Method*, Dorsey Press, Belmont (CA).

Small, T. A. (2011) 'What the Hashtag? A Content Analysis of Canadian Politics on Twitter', *Information, Communication & Society*, 14(6): 872-895.

Thaller, M. (1991) 'The Historical Workstation Project', *Computers and the Humanities*, 25(2): 149-162.

Thomas III, W. G. (2004) 'Computing and the Historical Imagination', in Schreibman, S., Siemens, R., & Unsworth, J. (eds) *A Companion to Digital Humanities*, Blackwell Publishing, Oxford: 56-68.

Wilkens, M. (2013) 'The Geographic Imagination of Civil War-Era American Fiction', *American Literary History*, 25(4): 803-840.

[1] It is also important to acknowledge that reactions to postmodern approaches are present as well in the historiographic debate (see for example Evans, 2001).

[2] See for example the adoption of social science methodologies in historical research in Fogel and Engerman (1974).

[3] However, the relationship between history and computing on the one side and literary and linguistic computing on the other side has always been complicated (see for example Robertson, 2016).

[4] As described in the FAQ section of the Internet Archive, a website owner can request to stop crawling or archiving a site and the Internet Archive will endeavor to comply to it. This will be signaled by a 'blocked site error' message such as 'This URL has been excluded from the Wayback Machine'.

[5] In 2001 the University of Bologna website won the 'WWW' prize from the Italian economic newspaper *Il Sole 24 Ore* for the best website in the category 'School, university and research'. Then, for three consecutive years (2005-2007) Unibo.it received the 'Osc@r del web' prize as the best Italian public administration website. In 2007 Luigi Nicolais, the Italian Minister of Public Administration, was also present to confer the prize.